COMPLETE DONOR-SET GUIDANCE FOR FROZEN DIFFUSION DOCKING MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Coordination information can guide a pretrained docking model, but its value depends on how specifically donor atoms are identified and how that information is used. We study a common Complete-Identity Contract Guidance (CICG) interface that applies a donor set during the final two denoising steps without retraining the generator or its confidence model. Historical Contract Transfer (HCT) supplies reference-mapped atom identities; trajectory-specific nearest-donor policies instead use historical donor counts, with or without elemental quotas, to select atoms from intermediate poses. No reference ligand coordinates are used as pose targets. We evaluate these policies in a pooled follow-up analysis of 86 assessable queries in 33 structural groups. Across two seeds, historical guidance raises joint pose, coordination, and validity success from 21.51% to 36.63%: +15.12 percentage points (95% group-bootstrap interval, 8.00–24.63). Composition-matched nearest-donor guidance reaches 33.14%, leaving the historical advantage unresolved at +3.49 points [-0.42, 8.87]. A retrospective audit using native donor labels finds that all 57 successes of the composition-matched policy occur on selected trajectories whose assigned sets match the native donor set, while 50 other exact assignments fail. Under the same donor-aware selector, a post-hoc comparison finds an additional 7.56-point gain [1.11, 14.04] from coordinate guidance, accompanied by lower ligand accuracy. These results support the usefulness of several donor-information policies within the evaluated scope, rather than a general advantage of fixed historical membership.

1 INTRODUCTION

A ligand binding a metalloprotein must satisfy two coupled requirements: its overall conformation must fit the protein pocket, and appropriate ligand atoms must coordinate the metal. Position, orientation, and torsion angles are continuous variables, whereas the identities of coordinating atoms form a discrete set. Two possible sets can share both donor count and elemental composition yet constrain different parts of the same molecule. For example, $\{O_a, O_b\}$ and $\{O_a, O_c\}$ each contain two oxygen donors but specify different atom memberships.

Pretrained diffusion models generate ligand poses through learned protein–ligand representations and geometric sampling (Corso et al., 2023; Cao et al., 2025). In common preprocessing pipelines, metal ions may be removed along with other receptor heteroatoms. Even when a metal ion is retained, general-purpose diffusion docking models do not explicitly condition pose generation on the complete set of ligand atoms assigned to coordinate it. Consequently, a pose can fit the protein pocket while realizing an incorrect metal–donor relationship. External coordination information offers a way to modify their predictions without retraining. In analogue design, a known metal-binding group may be retained while peripheral substituents change (Bonardi et al., 2020). A donor set then provides a partial condition while leaving the remaining pose to the generator. However, a binding mode observed for one compound need not transfer correctly to another, and less specific information may still be useful.

We study how the specificity and use of donor information affect docking predictions through a common inference-time interface. **Complete-Identity Contract Guidance (CICG)** holds a supplied donor set fixed during the last two denoising steps and reduces metal–donor distance residuals through ligand translation, rotation, and torsion updates. The generator and confidence model retain their

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

pretrained parameters. A *contract* denotes the requested complete ligand-side donor membership, not a guarantee that the generated pose realizes that membership. **Historical Contract Transfer (HCT)** supplies sets from earlier complexes at compatible receptor sites, without using the query's native donor labels or copying reference ligand coordinates into pose targets.

Historical transfer and nearest-donor selection construct alternative inputs to the same CICG operator. The nearest-donor policies retain historical cardinality, optionally with elemental composition, and choose atoms separately for each trajectory. Native-donor guidance provides a diagnostic comparison, while a post-hoc historical-donor reranker tests selection without coordinate modification. These comparisons examine how information specificity and its use affect docking outcomes.

We evaluate alternative donor-set construction policies under a common frozen-model guidance interface and matched sampling conditions. A pooled follow-up on 86 assessable queries shows improved joint Top-1 success with both historical and trajectory-local contracts, while their relative advantage remains uncertain. A post-hoc bank-by-selector comparison further identifies additional joint utility from coordinate guidance under the same donor-aware rule, together with losses in ligand accuracy. These analyses distinguish contract assignment, geometric realization, and final selection.

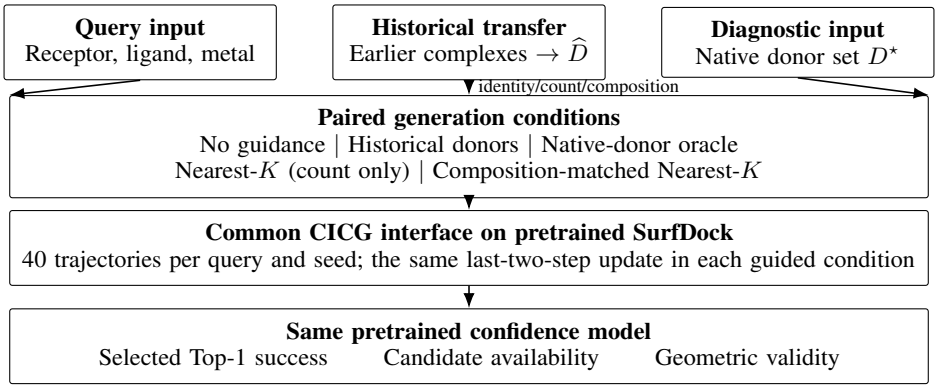


Figure 1: CICG with alternative donor-set construction policies. HCT supplies fixed identities or the counts and elemental quotas used by nearest-donor selection. Guided conditions share the correction operator, pretrained models, sampling budget, and corresponding noise. The post-hoc donor-aware selector is applied to both baseline and historical-guided banks (Section 3.5).

2 Related work

Docking with metal and structural information. DiffDock models ligand translation, rotation, and torsions with diffusion; SurfDock incorporates protein-surface features into this geometric framework (Corso et al., 2023; Cao et al., 2025). Metal-specific approaches incorporate coordination knowledge through different mechanisms. AutoDock4Zn modifies a docking potential (Santos-Martins et al., 2014). GM-DockZn generates poses by matching potential coordinating atoms to zinc coordination geometries. It also evaluates coordination numbers and coordinating-atom identities alongside pose RMSD, recognizing that a low-RMSD pose can still have incorrect metal coordination (Wang et al., 2020). MetalloDock learns metal-aware ligand generation (Zhang et al., 2026). Constrained and template docking use known structural relationships, and AutoDock Bias incorporates atom-specific interaction information into docking scores (Allen et al., 2015; Arcon et al., 2019). Metalloprotein Bias Docking (MBD) extends AutoDock Bias with knowledge-derived metal-interaction biases and was evaluated across Ca, Co, Fe, Mg, Mn, and Zn (Clemente et al., 2024). HCT uses reference structures to establish site compatibility and transfer donor identities, while CICG applies those identities to a pretrained generator.

Guidance of pretrained diffusion models. GeoDirDock guides translational, rotational, and torsional variables and supports template-based docking through angle transfer on a maximum common substructure (Miñán et al., 2025). HCT transfers donor membership without using reference ligand coordinates or torsions as pose targets. Universal Guidance uses external objectives with

pretrained diffusion models, while Projected Diffusion incorporates constraints during sampling (Bansal et al., 2024; Christopher et al., 2024). CICG uses local kinematic correction to compare donor-set construction policies within a common pretrained generator.

Generation, selection, and validity. Because correction changes both candidate geometry and confidence-model inputs, we distinguish candidate availability from selected-pose success. PoseBusters emphasizes that low root-mean-square deviation (RMSD) alone does not ensure physical validity (Buttenschoen et al., 2024). Interaction-recovery studies further distinguish pose placement from recovery of protein contacts (Errington et al., 2025). Our coordination criterion tests a complete ligand donor set and its geometry under one whole-ligand mapping. More generally, prediction accuracy and downstream decision quality need not coincide (Elmachtoub & Grigas, 2022; Mandi et al., 2022). We examine this distinction for donor-set correctness, selected docking success, and candidate-level clashes using a fixed confidence model.

3 Complete-Identity Contract Guidance

3.1 A donor-set condition for a frozen generator

A query is $X = (P, L, s_m, \mathbf{z}_m)$, consisting of the available receptor P, chemically specified ligand graph L, metal element s_m, and metal coordinate $\mathbf{z}_m$. A donor contract $D \subseteq V(L)$ identifies the ligand atoms intended to coordinate that metal. Completeness refers to all ligand atoms assigned to coordinate the selected metal. Atom indices are mapped to the query ligand before sampling.

Let $\mathcal{B}_0 = G_\theta(X;\xi)$ be the unmodified candidate bank and $\mathcal{B}_D = \mathcal{A}(G_\theta, X, D;\xi)$ the donor-guided bank. Here ξ denotes the initial randomness and subsequent trajectory noise. The same pretrained generator parameters θ and confidence parameters η are used in every condition. For each trajectory, $D_t = D$ throughout its intervention. Historical and native-donor policies share one set across trajectories; nearest-donor policies construct a set separately for each trajectory. All use the same CICG update, while coordinates remain free to change through the model's kinematic parameterization.

3.2 Coordinate update

At an intervention, let $q = (\mathbf{t}, \boldsymbol{\omega}, \varphi_1, \ldots, \varphi_{n_{\mathrm{rot}}})$ denote local translation in Å and rotation-vector and rotatable-bond coordinates in radians. The backbone's coordinate-update routine gives atom positions $\mathbf{x}_i(q)$. We minimize a radial residual on the supplied donors,

$$r_i(q) = \frac{\|\mathbf{x}_i(q) - \mathbf{z}_m\|_2 - d^{\mathrm{tar}}_{s_m,e_i}}{\sigma}, \qquad i \in D, \qquad \sigma = 0.10\,\text{Å}, \tag{1}$$

where e_i is the donor element. The metal–element targets are fixed constants (Appendix A), giving a radial-only objective on the supplied donors.

At iteration j, a central finite-difference Jacobian J_j is computed with step 10^{-5} in the stored kinematic coordinates. With $M = \mathrm{diag}(1, 1, 1, 1, 1, 1, 8, \ldots, 8)$, the update is

$$\Delta q_j = -M^{-1/2}\big(J_j M^{-1/2}\big)^{+} r(q_j). \tag{2}$$

The pseudoinverse is calculated by singular value decomposition (SVD), with relative cutoff 10^{-10}. This gives the minimum-metric-norm least-squares correction to the linearized residual. The nonlinear problem is addressed by relinearizing after accepted updates.

We try scales $\alpha \in \{1, 2^{-1}, \ldots, 2^{-15}\}$. A trial is accepted only if its coordinates are finite, its residual norm decreases, and its heavy-atom Cartesian RMS displacement from the current iterate does not exceed 0.25 Å. There are at most five relinearizations per intervention, with early stopping at $\|r\|_2 \le 10^{-3}$. An unsuccessful trial leaves the current iterate unchanged. The operator is applied after steps 18 and 19 of the 20-step SurfDock sampler, using zero-based indexing, to all 40 trajectories. The displacement cap applies to each accepted update; receptor clashes are evaluated after sampling.

3.3 HISTORICAL CONTRACT TRANSFER

HCT uses a collection $\mathcal{H}$ of earlier complexes to return one query donor set or abstain:

$$\widehat{D} = S_H(X, \mathcal{H}) \in \mathcal{C}(L) \cup \{\perp\}, \tag{3}$$

where $\mathcal{C}(L)$ is the set of admissible ligand donor sets. The reference collection contains 2,344 usable records retained from 2,423 raw records, released before 2023 and before the corresponding query. The strict matching rule requires the same metal; at least 90% sequence identity and 90% coverage in each alignment direction; matching aligned protein first shells defined within 3 Å of the metal; an earlier reference release; different ligand components and canonical graphs; and agreement of mapped donor sets.

Two extensions increase coverage. Donor-core matching replaces a global maximum common substructure with the radius-two graph neighborhood of the reference donor atoms. Graph-deduplicated consensus counts repeated observations with the same canonical ligand graph as a single vote. These rules and their outputs were fixed before comparison with query-native donor labels. The historical coordinates are used to identify reference donors and compare receptor sites, but only the mapped donor identities enter the correction. Appendix A.2 details the resulting coverage and precision.

On queries receiving a contract, the guided bank replaces the unmodified bank and the original confidence model selects the final pose:

$$\Pi(X) = \begin{cases} \mathrm{Sel}_\eta(\mathcal{B}_{\widehat{D}}), & \widehat{D} \neq \perp, \\ \mathrm{Sel}_\eta(\mathcal{B}_0), & \widehat{D} = \perp, \end{cases} \qquad \mathrm{Sel}_\eta(\mathcal{B}) = \arg\max_{y \in \mathcal{B}} s_\eta(y). \tag{4}$$

All supplied contracts are applied, and confidence alone selects the returned pose.

3.4 TRAJECTORY-SPECIFIC NEAREST-DONOR POLICIES

Nearest-K uses the historical cardinality $K_H = |\widehat{D}|$ but selects one set independently for each trajectory immediately before its step-18 correction. Let $\mathbf{x}^{(18)}$ be that pose and $V_{\mathrm{eligible}}(L, s_m)$ the supported nitrogen, oxygen, and sulfur atoms. We select

$$D_{\mathrm{near}} = \arg \min_{D \subseteq V_{\mathrm{eligible}},\, |D| = K_H} \sum_{i \in D} \|\mathbf{x}_i^{(18)} - \mathbf{z}_m\|_2. \tag{5}$$

Distance ties are resolved by input atom index. Eligibility is oxygen for calcium; nitrogen or oxygen for magnesium; and nitrogen, oxygen, or sulfur for cobalt, iron, manganese, and zinc. The chosen set is held fixed at step 19 and throughout all finite-difference, line-search, and relinearization calculations. Historical information still supplies both K_H and the applicable query set.

The *composition-matched Nearest-K* control additionally preserves the historical count $n_e(\widehat{D})$ of each element $e \in \{\mathrm{N}, \mathrm{O}, \mathrm{S}\}$. Writing V_e for eligible atoms of element e, it selects

$$D_{\mathrm{comp}} = \arg \min_{D \subseteq V_{\mathrm{eligible}}} \sum_{i \in D} \|\mathbf{x}_i^{(18)} - \mathbf{z}_m\|_2 \quad \text{subject to} \quad |D \cap V_e| = n_e(\widehat{D}) \text{ for every } e. \tag{6}$$

This holds the target-distance composition fixed while changing which atoms receive the constraints. Selection occurs once per trajectory at step 18, and membership remains fixed thereafter. Both nearest-donor controls inherit historical scope and cardinality; the composition-matched control also inherits the potentially incorrect elemental counts.

3.5 DONOR-AWARE SELECTION AND COORDINATE-BANK ABLATION

We apply one deterministic donor-aware rule to both the unmodified and historical-guided 40-pose banks to compare coordinate construction under a common selection policy. For a pose y, let $N_{K_H}(y)$ be its nearest K_H ligand heavy atoms and G_L the chirality-preserving automorphisms of the input ligand graph. The overlap deficit is

$$e(y) = K_H - \max_{g \in \mathcal{G}_L} |N_{K_H}(y) \cap g(\widehat{D})|. \tag{7}$$

Candidates are ordered by increasing $e(y)$, decreasing frozen-model confidence for that pose, and increasing sample identifier. Selection uses the input ligand graph, historical donor set, and evaluated bank; native labels enter evaluation only. The unchanged rule is applied separately to both banks. The reranking and bank-by-selector comparisons were specified after inspection of the original experiments and are post-hoc analyses.

4 EXPERIMENTAL DESIGN

Data and analysis population. The initial roster contains 304 queries: 179 non-overlapping entries from a 400-complex collection and 125 from a separate post-2023 collection. HCT supplies 62 queries in 17 structural/site groups, giving 20.39% coverage. An additional 38 queries in 22 groups were selected from screened Protein Data Bank collections (Berman et al., 2000) with fixed metal quotas, without query-native donor labels or docking outcomes. Their structure entries and frozen group identifiers do not overlap the initial cohort. Appendix A.2 gives the selection rules, release ranges, and metal composition.

The stereochemistry checker's input domain covers specified carbon tetrahedral centers and E/Z bonds. An input-only mask excludes 14 ligands with specified non-carbon tetrahedral centers or ligand-internal metals, leaving 86 queries in 33 groups (55 initial and 31 additional). All assessable queries are retained across conditions and seeds. We adopted this pooled, domain-restricted analysis during reporting, after the original outcomes were known. Appendix B.1 reports the original cohort-specific comparisons, including the 62- and 100-query analyses that counted unsupported inputs as failures.

Information and matched execution. Each query is sampled at seeds 42 and 43 with 40 trajectories per condition. Historical transfer was fixed before its native-label audit. Native labels subsequently define evaluation and retrospective correctness strata; they do not select which historical contracts are applied. The nearest-donor controls inherit historical cardinality, and the composition-matched control also inherits its elemental quotas.

Inputs, generator and confidence weights, initial conditions, and corresponding sampling noise are matched. Each guided arm uses the same radial targets, update operator, correction steps, and maximum iteration count. The native-donor oracle supplies atom identities, not native ligand coordinates, distances, or angles as correction targets. Sample counts and model steps are matched; actual correction runtime can vary. Replication, nearest-donor comparisons, and deterministic reranking were sequential follow-ups to the initial historical-guidance experiment; reranking and label-conditioned diagnostics are post-hoc.

Joint success. We call the evaluation's *clash-free Robust Full* endpoint *joint success*. It requires symmetry-aware heavy-atom root-mean-square deviation (RMSD) below $2.0\,\text{Å}$ in the receptor frame and a coordination witness: one chirality-preserving whole-ligand mapping under which the $K_\star = |D^\star|$ nearest ligand heavy atoms are the native donors, each radial error is at most $0.50\,\text{Å}$, and each metal-to-donor vector differs from its native direction by at most 30°. All heavy-atom elements enter the evaluation's nearest-atom search. Ligand RMSD is minimized separately from the coordination witness.

The same pose must also have finite coordinates, successful molecular sanitization, preserved supported input stereochemistry, and no ligand–receptor heavy-atom separation below $1.2\,\text{Å}$. These are evaluation criteria, not generation targets. Appendix A.3 defines the geometry-valid flag and distinguishes these checks from the full PoseBusters battery. Assigned-set exactness is a separate diagnostic: it asks whether the set fixed at the intervention agrees with the native donor annotation, rather than whether the final coordinates realize that set.

Selection and statistics. Top-k success means that at least one of the first k poses in a method's output ranking passes the joint criterion. Sampling conditions use frozen confidence, whereas donor-aware reranking uses Equation (7). Top-40 equals availability for a 40-pose bank. Let R^a_{is} be Top-1

success for condition a, query i, and seed s. The paired effect is

$$d_i^{a,b} = \frac{1}{2} \sum_{s \in \{42,43\}} (R_{is}^a - R_{is}^b), \qquad \widehat{\Delta}^{a,b} = \frac{1}{86} \sum_{i=1}^{86} d_i^{a,b}. \tag{8}$$

Baseline contrasts assess the benefit of guidance, while historical-versus-nearest contrasts quantify the incremental value of fixed historical membership. Group bootstrap intervals retain all query and seed pairs within a resampled group. The original four-condition and reranking analyses use 20,000 common draws within the assessable cohort; the composition-matched comparisons come from the completed follow-up analysis. We report nominal two-sided 95% intervals. Secondary and retrospective contrasts are not multiplicity-adjusted, and an interval containing zero establishes neither superiority nor equivalence. The 172 query–seed outcomes are descriptive observations, not 172 independent biological samples. Appendix A.4 preserves analysis chronology and the earlier resampling settings.

5 Results

5.1 Donor guidance improves selected coordination-aware predictions

Historical guidance increases joint Top-1 success from 37/172 to 63/172, or 21.51% to 36.63%, in the pooled follow-up analysis (Table 1). The paired gain is 15.12 percentage points (95% interval, 8.00–24.63; Table 2), with 13 additional successes at each seed. Native-donor guidance achieves 70/172 successes, a 19.19-point gain [11.21, 28.51] over no guidance.

Table 1: Joint Top-1 success in the pooled follow-up (86 queries, 33 groups, two seeds). Counts summarize 172 query–seed outcomes from 40-pose banks. Donor-aware reranking is a post-hoc selection-only comparison.

Condition	Successful query–seed outcomes	Mean success (%)
No guidance	37/172	21.51
Nearest-K (count only)	56/172	32.56
Composition-matched Nearest-K	57/172	33.14
Historical donors	63/172	36.63
Native-donor oracle	70/172	40.70
Donor-aware reranking	51/172	29.65

Table 2: Paired joint Top-1 differences in percentage points. Effects average two seeds within each query; 95% intervals resample structural/site groups. Reranking comparisons are post-hoc. Bootstrap settings are given in Appendix A.4.

Comparison	Difference	95% confidence interval
Historical donors − no guidance	+15.12	[8.00, 24.63]
Nearest-K (count only) − no guidance	+11.05	[5.49, 17.86]
Composition-matched Nearest-K − no guidance	+11.63	[6.02, 18.94]
Native-donor oracle − no guidance	+19.19	[11.21, 28.51]
Historical − Nearest-K (count only)	+4.07	[0.00, 10.37]
Historical − composition-matched	+3.49	[−0.42, 8.87]
Native-donor oracle − composition-matched	+7.56	[1.52, 14.42]
Donor-aware reranking − no guidance	+8.14	[2.73, 15.79]
Historical guidance − donor-aware reranking	+6.98	[1.28, 12.86]

5.2 Nearest-donor policies recover much of the aggregate benefit

Count-only and composition-matched Nearest-K achieve 32.56% and 33.14% joint success, respectively, recovering much of the improvement from historical guidance. Historical membership adds

4.07 points [0.00, 10.37] over the count-only policy and 3.49 points [-0.42, 8.87] over the composition-matched policy (Table 2); its overall advantage remains unresolved. Native-donor guidance exceeds the composition-matched policy by 7.56 points [1.52, 14.42], although this oracle can also correct historical cardinality and elemental-quota errors. Nearest-donor selection is therefore a competitive use of partial historical information.

5.3 Assignment exactness, realization, and selection are distinct

A retrospective native-label audit finds exact sets in 39.07% of composition-matched trajectories and 62.21% of that policy's confidence-selected trajectories (Table 3). Among 107 selected trajectories with exact assignments, 57 pass the joint criterion; all 65 with incorrect assignments fail. The remaining 50 failures with exact assignments show that contract accuracy alone is insufficient. These counts describe an association under an endpoint that itself requires native coordination recovery.

Table 3: Retrospective assignment diagnostics on 86 queries. Exactness compares the pre-correction donor set with the native annotation; selected assignments belong to the composition-matched policy's confidence-selected trajectory. Rates are descriptive.

Diagnostic	Count	Rate (%)
Native-exact assignments across all trajectories	2,688/6,880	39.07
Native-exact assignment on the selected trajectory	107/172	62.21
Queries with at least one exact assignment across both seeds	78/86	90.70
Joint success given an exact selected assignment	57/107	53.27
Joint success given an incorrect selected assignment	0/65	0.00

Historical sets are exact on 69 assessable queries and incorrect on 17. Relative to no guidance, the respective paired effects are $+21.74$ points [12.70, 33.13] and -11.76 points [-20.00, 0.00]. Relative to count-only Nearest-K, they are $+7.97$ [1.39, 16.28] and -11.76 [-23.81, 3.90]. These retrospectively defined strata reveal sensitivity to transfer errors.

5.4 Coordinate guidance adds benefit under a fixed selector

We apply the same donor-aware rule to both the unmodified and historical-guided banks in a post-hoc 2×2 comparison (Table 4). Under donor-aware selection, historical guidance increases joint success from 51/172 to 64/172: +7.56 percentage points (95% group-bootstrap interval, 1.11–14.04), with 21 improvements and eight regressions. The gains are positive at both fixed seeds. Applying donor-aware selection to the guided bank adds only 0.58 points [-1.75, 3.37] over its original confidence ranking.

Under the same donor-aware selector, ligand success decreases from 107/172 to 99/172 (-4.65 points [-8.06, -0.68]), and selected clashes increase from 12/172 to 21/172 (+5.23 points [-0.71, 11.21]). Thus changing the coordinate bank provides additional joint utility under a fixed selection rule, with a trade-off in ligand accuracy. This comparison includes the downstream changes in both geometry and frozen-model confidence.

5.5 Coordination recovery can sacrifice ligand accuracy

We also hold the baseline outcome fixed to examine recoveries. Among 69 query–seed outcomes with correct baseline ligand placement but failed joint success, historical guidance recovers joint success in 26 and loses ligand accuracy in 14. The narrower ligand-correct/identity-wrong subset contains 53 outcomes, with 18 joint recoveries and 12 ligand losses. Both subsets are defined from baseline outcomes (Appendix B.12). At Top-40, historical guidance succeeds in 96/172 banks versus 106/172 for composition-matched Nearest-K, despite its higher Top-1 count. Of the 33 historical Top-1 improvements over no guidance, 25 occur where the baseline bank already contains a joint-success pose. The selected-pose gain therefore combines changes in candidate realization and selection.

Historical guidance increases candidate clashes and lowers bank-wide geometry validity (Table 5). Its selected ligand-only successes decrease from 106 to 101. Both nearest-donor policies also worsen bank-wide validity. Thus a gain in the clash-free joint endpoint coexists with degradation elsewhere in

Table 4: Post-hoc bank-by-selector comparison on 86 queries in 33 groups. Entries count joint Top-1 successes across 172 query–seed observations. Each bank contains 40 poses; both donor-aware conditions use the same transferred set and selection rule.

Coordinate bank	Frozen confidence	Donor-aware selection
No guidance	37/172	51/172
Historical guidance	63/172	64/172

the candidate population. The radial objective has no receptor-clash term, and its local displacement cap does not prevent collisions.

Table 5: Geometric outcomes for five sampling conditions on 86 queries. Selected counts summarize 172 query–seed outcomes; candidate fractions use 6,880 poses per condition. Composition-matched physical results for the initial 55-query cohort appear in Appendix B.11.

Condition	Ligand Top-1	Top-1 clashes	All-pose clash (%)	Geometry-valid (%)
No guidance	106	11	16.83	83.17
Historical donors	101	22	27.22	72.78
Nearest-K (count only)	109	20	26.32	73.68
Native-donor oracle	101	21	27.17	72.83
Composition-matched Nearest-K	105	22	29.39	70.61

Appendix F illustrates a recovery and a regression. A separate native-donor experiment with SurfDock and DiffDock-L (v1.1 checkpoints) appears in Appendix C.1.

6 Discussion

CICG supplies a common control interface under which historical and trajectory-local donor policies both improve on unmodified SurfDock. The central question is therefore not only whether donor information helps, but how it should be used. Historical atom assignments improve selected coordination-aware predictions, yet nearest-donor guidance obtains much of the benefit using only donor count and elemental quotas. The observed overall advantage of fixed historical identity remains unresolved. These alternatives differ in information specificity: a fixed set conveys a stronger instruction when correct, whereas trajectory-specific selection can avoid some transferred membership errors without abandoning all historical information.

The assignment audit separates contract construction, geometric realization, and final selection. Exact assignments were common among successful outputs, but many exact assignments still failed, underscoring the importance of downstream geometry and ranking.

Our comparisons hold the correction operator fixed and examine information-use policies. They do not isolate interleaved guidance from applying the same operator exclusively after sampling. Reranking demonstrates that some benefit is obtainable without changing coordinates. The lower validity of guided banks also shows that more accurate coordination and broader physical quality are separate objectives.

Scope and statistical interpretation. The 86-query pooled follow-up analysis combines the two collected cohorts and applies the checker's input-domain rule. This reporting-stage choice changes the denominator, not the existing successful predictions; the original cohort analyses remain available. The initial 62/304 coverage estimate does not describe the added cohort or all metalloproteins. Full-roster fallback gains cannot be inferred by applying the pooled 86-query effect to the original coverage.

Metal quotas broaden the collected sample, but only three magnesium queries remain assessable, alongside 22 zinc queries. Query and group non-overlap does not guarantee protein-family independence, and two fixed seeds do not cover all sampling variability. Group-deletion results in

Appendix B.10 leave the original analyses unchanged and are descriptive sensitivity checks. Initial-cohort concentration of the composition-matched contrast is not a group-level analysis of the expanded cohort. Chronological reference exclusion also does not eliminate pretrained-model exposure.

Water occupancy, protonation, receptor rearrangement, and electronic states are not separately resolved. The method requires access to the sampler's coordinate-update interface. Equal-information comparisons with coordinate templates, metal-specific docking systems, and general guidance algorithms remain outside this evaluation. Historical transfer to a second backbone has not been established by the separate native-donor experiment.

7 Conclusion

CICG enables pretrained docking models to use donor information through fixed historical assignments or trajectory-specific nearest-donor selection. In the pooled follow-up analysis, both policies improve selected joint success over unmodified SurfDock, while their relative performance remains uncertain. A post-hoc comparison under the same donor-aware selector also supports additional joint utility from coordinate guidance, with a loss in ligand accuracy. Assignment errors and increased receptor clashes remain important limitations. Together, these findings support evaluating contract construction, coordinate realization, and final selection jointly when incorporating partial coordination information into frozen docking models.

REPRODUCIBILITY STATEMENT

We provide an anonymized supplementary archive containing 800 stored pose banks (32,000 poses) from the four-condition, 100-query analysis and 172 composition-matched banks (6,880 poses) for the 86-query assessable follow-up. The archive includes ligand coordinates, confidence scores and sample identifiers, query and structural-group assignments, transferred donor sets and provenance, intervention-step assignment traces, and pose- and query-level evaluation outcomes. Executable CPU scripts reproduce the fixed-bank evaluation, donor-aware reranking, assessable-domain analyses, composition-matched endpoint calculations, and paired group-bootstrap comparisons from the stored outputs. Input ordering, resampling settings, environment specifications, and file checksums are included. Appendix A.5 describes the software environment and reproduction commands. The supplement supports analysis from saved coordinates and confidence scores; end-to-end regeneration of the original diffusion sampling additionally requires the pretrained model checkpoints and upstream preprocessing assets described in Appendix A.5.

ETHICS STATEMENT

This study uses computational molecular-structure data and involves no human participants or patient data. Docking and geometric validity do not establish biological activity, toxicity, or clinical efficacy. Structures, software, and pretrained models remain subject to their respective licenses.

AI USE STATEMENT

The authors set the research scope, made the methodological and experimental decisions, directed the computational work, and determined the interpretation and presentation of the findings. Large language model tools, including ChatGPT and Codex, assisted literature discovery and synthesis, hypothesis and experimental-design discussions, method development, implementation and debugging, data organization, mathematical reasoning and exposition, analysis and interpretation of computational outputs, visualization of saved molecular coordinates, and manuscript drafting and revision. The authors selected and refined the suggestions used in the study. Validation included input and atom-mapping checks, matched-condition comparisons, coordinate and endpoint consistency tests, and replay of the reported statistics from saved outputs. All empirical findings are based on the computational experiments and analyses described in this paper. The authors retain responsibility for the scientific decisions, results, citations, and final manuscript.

REFERENCES

William J. Allen, Trent E. Balius, Sudipto Mukherjee, Scott R. Brozell, Demetri T. Moustakas, P. Therese Lang, David A. Case, Irwin D. Kuntz, and Robert C. Rizzo. DOCK 6: Impact of new features and current docking performance. *Journal of Computational Chemistry*, 36(15): 1132–1156, 2015. doi: 10.1002/jcc.23905.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025. doi: 10.1214/24-AOAS1998.

Juan Pablo Arcon, Carlos P. Modenutti, Demian Avendaño, Elias D. Lopez, Lucas A. Defelipe, Francesca Alessandra Ambrosio, Adrian G. Turjanski, Stefano Forli, and Marcelo A. Marti. AutoDock Bias: improving binding mode prediction and virtual screening using known protein–ligand interactions. *Bioinformatics*, 35(19):3836–3838, 2019. doi: 10.1093/bioinformatics/btz152.

Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Roni Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=pzpWBbnwiJ.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):43:1–43:34, 2021. doi: 10.1145/3478535.

Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 2000. doi: 10.1093/nar/28.1.235.

Alessandro Bonardi, Alessio Nocentini, Silvia Bua, Jacob Combs, Carrie Lomelino, Jacob Andring, Laura Lucarini, Silvia Sgambellone, Emanuela Masini, Robert McKenna, Paola Gratteri, and Claudiu T. Supuran. Sulfonamide inhibitors of human carbonic anhydrases designed through a three-tails approach: Improving ligand/isoform matching and selectivity of action. *Journal of Medicinal Chemistry*, 63(13):7422–7444, 2020. doi: 10.1021/acs.jmedchem.0c00733.

Martin Buttenschoen, Garrett M. Morris, and Charlotte M. Deane. PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chemical Science*, 15(9):3130–3139, 2024. doi: 10.1039/D3SC04185A.

Duanhua Cao, Mingan Chen, Runze Zhang, Zhaokun Wang, Manlin Huang, Jie Yu, Xinyu Jiang, Zhehuan Fan, Wei Zhang, Hao Zhou, Xutong Li, Zunyun Fu, Sulin Zhang, and Mingyue Zheng. SurfDock is a surface-informed diffusion generative model for reliable and accurate protein–ligand complex prediction. *Nature Methods*, 22(2):310–322, 2025. doi: 10.1038/s41592-024-02516-y.

Jacob K. Christopher, Stephen Baek, and Ferdinando Fioretto. Constrained synthesis with projected diffusion models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/a28af221f2f70be183afc16797a56b91-Abstract-Conference.html.

Camila M. Clemente, Juan M. Prieto, and Marcelo Martí. Unlocking precision docking for metalloproteins. *Journal of Chemical Information and Modeling*, 64(5):1581–1592, 2024. doi: 10.1021/acs.jcim.3c01853.

Gabriele Corso, Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi Jaakkola. DiffDock: Diffusion steps, twists, and turns for molecular docking. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.01776.

Gabriele Corso, Arthur Deng, Nicholas Polizzi, Regina Barzilay, and Tommi Jaakkola. Deep confident steps to new pockets: Strategies for docking generalization. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/db334db287337b2a365120b524300ef3-Abstract-Conference.html.

Adam N. Elmachtoub and Paul Grigas. Smart "predict, then optimize". *Management Science*, 68(1):9–26, 2022. doi: 10.1287/mnsc.2020.3922.

David Errington, Constantin Schneider, Cédric Bouysset, and Frédéric A. Dreyer. Assessing interaction recovery of predicted protein-ligand poses. *Journal of Cheminformatics*, 17(1):76, 2025. doi: 10.1186/s13321-025-01011-6.

Johann Gasteiger and Mario Marsili. Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges. *Tetrahedron*, 36(22):3219–3228, 1980. doi: 10.1016/0040-4020(80)80168-2.

Wolfgang Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 32(5):922–923, 1976. doi: 10.1107/S0567739476001873.

Greg Landrum, Paolo Tosco, Brian Kelley, et al. rdkit/rdkit: 2023_03_1 (Q1 2023) Release, 2023. URL https://zenodo.org/records/7880616.

Greg Landrum, Paolo Tosco, Brian Kelley, et al. rdkit/rdkit: 2025_09_6 (Q3 2025) Release, 2026. URL https://zenodo.org/records/18797641.

Jayanta Mandi, Víctor Bucarey, Maxime Mulamba Ke Tchomba, and Tias Guns. Decision-focused learning: Through the lens of learning to rank. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 14935–14947, 2022. URL https://proceedings.mlr.press/v162/mandi22a.html.

Raúl Miñán, Javier Gallardo, Álvaro Ciudad, and Alexis Molina. Informed protein–ligand docking via geodesic guidance in translational, rotational and torsional spaces. *Nature Machine Intelligence*, 7:1555–1560, 2025. doi: 10.1038/s42256-025-01091-x.

Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. CatBoost: Unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://papers.neurips.cc/paper_files/paper/2018/hash/14491b756b3a51daac41c24863285549-Abstract.html.

Diogo Santos-Martins, Stefano Forli, Maria João Ramos, and Arthur J. Olson. AutoDock4Zn: An improved AutoDock force field for small-molecule docking to zinc metalloproteins. *Journal of Chemical Information and Modeling*, 54(8):2371–2379, 2014. doi: 10.1021/ci500209e.

Kai Wang, Nan Lyu, Hongjuan Diao, Shujuan Jin, Tao Zeng, Yaoqi Zhou, and Ruibo Wu. GM-DockZn: A geometry matching-based docking algorithm for zinc proteins. *Bioinformatics*, 36(13): 4004–4011, 2020. doi: 10.1093/bioinformatics/btaa292.

Hui Zhang, Xujun Zhang, Qun Su, Yangyang Zheng, Linlong Jiang, Kai Zhu, Qiaolin Gou, Odin Zhang, Shi Li, Bo Peng, Shaokai Ni, Yushen Du, Jiayi Tang, Yu Kang, Chang-Yu Hsieh, Dan Li, Wenteng Chen, Tingjun Hou, and Peichen Pan. MetalloDock: Decoding metalloprotein–ligand interactions via physics-aware deep learning for metalloprotein drug discovery. *Journal of the American Chemical Society*, 148(3):3086–3101, 2026. doi: 10.1021/jacs.5c15876.

A Implementation and evaluation details

A.1 Correction parameters

The correction operates in the pretrained sampler's native translation, rotation-vector, and torsion coordinates. At each of the final two reverse steps, it starts from the current sampled pose, computes the donor-distance residual and Jacobian, and accepts a backtracked update only when the residual decreases and the per-update Cartesian displacement satisfies the cap. Translations and atom positions are represented in Å, and rotation-vector and torsion components in radians. The finite-difference steps and metric weights refer to these coordinates. The numerical settings are listed in Table 6; the metal–element targets are given in Table 7.

Table 6: Numerical settings of the last-two-step correction. The finite-difference step and metric refer to the stored kinematic coordinates.

Quantity	Value
SurfDock reverse steps	20
Activated steps	After 18 and 19, zero-based indexing
Samples per query and seed	40
Coordinates	Translation, rotation vector, rotatable-bond torsions
Residual scale	0.10 Å
Jacobian	Central finite differences with step 10^{-5}
Metric diagonal	1 for translation/rotation; 8 for torsions
Linear solve	Metric-weighted SVD pseudoinverse
Relative singular-value cutoff	10^{-10}
Relinearizations per intervention	At most 5
Stopping tolerance	$\|r\|_2 \le 10^{-3}$
Backtracking scales	$\{2^{-b} : b = 0, \dots, 15\}$
Per-update heavy-atom RMS limit	0.25 Å
Acceptance	Finite coordinates, reduced residual norm, displacement limit
Clash and stereochemistry guard	None inside the correction
Angle, exclusion, or strain terms	None

Table 7: Fixed metal–donor-element target distances $d^{\mathrm{tar}}_{s_m,e}$ in Å. Dashes indicate unsupported pairs. These empirical targets are shared with the earlier seven-hook study; the correction algorithms and schedules differ.

Metal	N	O	S
Ca	—	2.40	—
Co	2.10	2.05	2.25
Fe	2.10	2.05	2.30
Mg	2.20	2.10	—
Mn	2.20	2.15	2.35
Zn	2.10	2.05	2.30

Equation (2) follows by writing $z = M^{1/2}\Delta q$. The linearized least-squares problem is $\min_z \|JM^{-1/2}z + r\|_2$. The minimum-Euclidean-norm solution in z is $-(JM^{-1/2})^{+}r$, yielding the minimum-M-norm correction in Δq. The rank cutoff and nonlinear line search define a local numerical correction. The displacement restriction is a Cartesian RMS cap on each accepted relinearization.

Kinematic mapping. Each correction call uses the sampled ligand as its reference conformer. Rigid rotation is about that conformer's centroid, followed by translation. Rotatable edges and their moving-atom masks are processed in the stored graph order; the first endpoint is stationary and the second belongs to the moving component. The flexible proposal is then Kabsch-aligned to the rigid proposal (Kabsch, 1976). Retraction adds translations, left-composes rotation matrices, and wraps accumulated torsions. The metric weights are one for each translation and rotation component and

eight for each torsion. The Cartesian displacement test permits a numerical tolerance of 10^{-12} Å. No collision or stereochemical test enters the line-search acceptance rule.

A.2 Historical data and transfer rules

The reference collection retains 2,344 usable records from 2,423 raw records. References are released before 1 January 2023 and strictly before the corresponding query. Candidates must have the same metal and a different ligand component and canonical isomeric molecular graph. The original 304-query roster combines 179 non-overlapping entries from a 400-complex collection with 125 entries from a separate post-2023 collection. Of these, 112 have a reference at the matched target/site and 62 receive a donor set. Native correctness is evaluated after transfer and does not select the replay population.

Receptor-site compatibility. A first shell consists of every standard protein residue with a non-hydrogen/deuterium atom within 3.0 Å of the metal. At least one reference/query chain pair must have global-alignment sequence identity and coverage in both directions of at least 90%. Alignment scores are one for a match, zero for a mismatch, -1 for gap opening, and -0.1 for extension. The first optimal alignment is used. The complete sets of aligned first-shell residue positions and residue identities must agree. The 3.0 Å threshold defines protein first-shell membership.

Atom transfer. Global maximum-common-substructure matching preserves elements, formal charges, exact bond orders, ring membership, and complete rings. It uses a five-second timeout and requires at least six atoms and 50% of the smaller ligand. At most 2,048 matches per ligand are enumerated. All reference donors must be included. Donor-local signatures compare element, formal charge, aromaticity, degree, and neighboring atom/bond states. A query receives a set only when the accepted projections agree.

The donor-centered extension instead matches atoms within two graph edges of the reference donors together with shortest paths connecting donor pairs. It uses an atom-mapped substructure query, ring/bond consistency, the same donor-local checks, and at most 4,096 matches. Multiple distinct projected donor sets are rejected rather than merged by evaluator symmetry. These transfer match calls do not explicitly enable chirality-sensitive matching, and no extra protonation or tautomer normalization is applied. This differs from the chirality-preserving whole-graph symmetry used in evaluation and in the selection-only comparator.

Reference consensus. Repeated references with the same canonical isomeric ligand graph contribute one vote only if they map successfully and agree. The graph-deduplicated winner must be unique, receive at least two graph votes and at least two-thirds of the votes, and remain the unique winner after any single graph vote is removed. The minimum-vote and vote-share thresholds are not reapplied inside this leave-one-vote-out check. Existing strict-rule assignments are retained; donor-centered assignments precede consensus assignments. Ties and nonunique mappings lead to abstention.

Table 8: Historical donor transfer on the original 304-query roster. Coverage is relative to all queries; precision is relative to supplied sets. Extension rows contain disjoint additions.

Transfer rule	Supplied	Correct	Coverage (%)	Precision (%)
Strict site and global-graph matching	33	25	10.86	75.76
Donor-centered core additions	13	9	—	69.23
Graph-deduplicated consensus additions	16	11	—	68.75
Combined transfer	62	45	20.39	72.58

The donor-centered extension supplies 13 of 61 previously unsupported queries. Consensus supplies 16 of 18 disagreement cases, all in one group. Combined transfer precision is 72.58% (17-group bootstrap interval [57.89, 84.75]%); 12 of its 17 errors retain the correct donor count. Reference donor names are mapped to indices only when the archived chemical-component heavy-atom name/element order matches the ligand graph. The supplement contains the execution contracts, candidate reference pairs, and reference metadata.

Additional-query selection. The additional cohort was selected from two previously screened structural collections: records through 15 July 2026 and a recent-release collection spanning 22 July–23 September 2026. Eligible records had to receive a historical contract, avoid Protein Data Bank entries already present in the initial cohort, and pass the recorded technical preparation criteria. An initial 60-query candidate set used square-root allocation by applicable metal counts, prioritizing recent releases and then a fixed hash order within metal. From this ordered set, fixed quotas selected 38 queries: three calcium, four cobalt, one iron, eight magnesium, one manganese, and 21 zinc complexes. Selection used neither native query-donor labels nor docking outcomes. These quotas deliberately broaden metal representation; they do not define a random sample of all metalloproteins.

Table 9: Metal composition of the evaluated historical-transfer cohorts. The combined sample includes the initial and additional queries.

Metal	Initial (62)	Additional (38)	Combined (100)
Calcium	14	3	17
Cobalt	21	4	25
Iron	9	1	10
Magnesium	2	8	10
Manganese	15	1	16
Zinc	1	21	22

The initial 62 and additional 38 queries have disjoint Protein Data Bank entries and no shared frozen group identifiers. Group construction uses receptor/site and reference connectivity for the strict-rule subset and sets of supporting reference identifiers for the extension and additional subsets. The resulting 17 and 22 groups are the dependence units retained in the analyses; distinct group identifiers do not establish protein-family independence. The combined 100-query sample has 39 groups, with 73 correct and 27 incorrect historical sets. The pooled follow-up applies the input-domain restriction described in Section 4. Coverage of 62/304 applies only to the original screened roster.

Table 10: Input-domain filtering by metal. Each entry is collected queries / assessable queries. Exclusion reflects the stereochemistry checker's supported representation, not chemical invalidity or generated-pose quality.

Metal	Initial	Additional	Combined
Calcium	14 / 8	3 / 2	17 / 10
Cobalt	21 / 21	4 / 4	25 / 25
Iron	9 / 9	1 / 1	10 / 10
Magnesium	2 / 1	8 / 2	10 / 3
Manganese	15 / 15	1 / 1	16 / 16
Zinc	1 / 1	21 / 21	22 / 22
Total	62 / 55	38 / 31	100 / 86

A.3 Pose-level success and stereochemistry

Native ligand coordinates $\mathbf{x}_i^\star$ and donor annotations $D^\star$ are fixed evaluation inputs in the receptor frame, separate from the historical-selector inputs. The archived annotations are mapped by ligand atom identity; the evaluation does not infer $D^\star$ from generated poses. Query labels and their atom-name/index correspondence are included in `evaluation/references.json` and the mapping provenance ledger. The release supports evaluation conditional on these annotations, but does not reconstruct the upstream chemical curation of every native donor label. This limits independent auditing of annotation ambiguity. With $K_\star = |D^\star|$, let $N_{K_\star}(y)$ be the nearest $K_\star$ ligand heavy atoms to the metal in pose y, including all heavy-atom elements and resolving distance ties by atom

index. A coordination witness is a chirality-preserving whole-ligand automorphism π satisfying

$$N_{K_\star}(y) = \pi(D^\star), \tag{9}$$

$$\max_{i \in D^\star} \left| \|\mathbf{x}_{\pi(i)}(y) - \mathbf{z}_m\|_2 - \|\mathbf{x}_i^\star - \mathbf{z}_m\|_2 \right| \leq 0.50\,\text{Å}, \tag{10}$$

$$\max_{i \in D^\star} \angle\big(\mathbf{x}_{\pi(i)}(y) - \mathbf{z}_m, \mathbf{x}_i^\star - \mathbf{z}_m\big) \leq 30^\circ. \tag{11}$$

All three conditions must hold under one mapping. Chemically cleaned, chirality-preserving full-graph mappings are enumerated with a cap of 10,000; cap exceedance fails the evaluation rather than truncating its witness search. Whole-ligand heavy-atom RMSD is minimized separately over the allowed mappings, without spatial alignment, and must be strictly below 2.0 Å. Its minimizing mapping need not be the coordination witness.

Stereochemistry is checked against the chemically specified input ligand. Each supported carbon tetrahedral center must retain the sign of its normalized signed triple product, whose magnitude is at least 0.02. For a specified E/Z double bond, input and output require $|\sin\tau| \leq 0.5$ and matching signs of $\cos\tau$. Specified non-carbon tetrahedral centers and ligand-internal metals are outside the checker's supported scope. The original full-cohort analyses count these inputs as failures. Joint-success tables in the main text pool the input-domain subset: 55 of 62 initial queries and 31 of 38 additional queries, totaling 86. This reporting-stage restriction retains every assessable query, regardless of its predicted validity or donor-set correctness. The 14 excluded input graphs and exclusion reasons are listed in `assessability/INPUT_DOMAIN_LEDGER.tsv`.

The geometry-valid flag is the conjunction of finite coordinates, successful RDKit sanitization (Landrum et al., 2026) of the input graph with predicted coordinates, preserved supported stereochemistry, and no fatal receptor clash. A fatal clash is any ligand/receptor heavy-atom separation strictly below 1.2 Å; nonfinite ligand coordinates also fail. The receptor coordinate reader retains non-hydrogen ATOM/HETATM records in the prepared receptor. Evaluation exceptions are recorded as failures, including a failed sanitization flag. These tests are narrower than the full PoseBusters battery (Buttenschoen et al., 2024). Joint success requires ligand success, a common coordination witness, and geometry validity.

Separate identity, radial, and directional diagnostics refer to the same selected pose. Each marginal tests existence over complete assignments, but their witnesses can differ; multiplying these marginal indicators would not reconstruct the joint criterion.

A.4 Paired statistical analysis

Query-level effects average the two fixed-seed paired differences before resampling. For each of 20,000 replicates, the bootstrap samples G of the G frozen group identifiers uniformly with replacement, concatenates all queries in each sampled group with their multiplicities, and averages the query effects. Groups are sampling units, not equally weighted group means. All conditions and both seeds stay paired. Two-sided intervals use the linearly interpolated 2.5th and 97.5th percentiles. Thus 124 or 200 query–seed outcomes do not constitute independent biological sample sizes.

The original analysis uses `numpy.random.default_rng`. For the initial cohort, pooled historical-minus-baseline, historical-minus-nearest, nearest-minus-baseline, and native-minus-baseline comparisons use random states 20272001, 20272002, 20272003, and 20272004, respectively. The expanded analysis uses 20263001, 20263003, 20263002, and 20263004 for the same contrasts. The post-hoc comparisons share state 20272001 for the initial cohort and 20263001 for the additional and combined cohorts. Per-seed and correctness-stratum random states are recorded in `METHODS_SETTINGS.json`.

Correctness strata are formed before calling the bootstrap function, and only groups represented in that stratum are sampled. All analyzed strata are nonempty. Contrast-specific random streams can produce slightly different Monte Carlo percentile endpoints for identical effect vectors; the original full-cohort intervals and streams are retained in Appendix B.1. The expanded historical-minus-nearest lower endpoint is exactly zero before rounding. The supplementary analysis reproduces the four central pooled historical-versus-baseline/nearest intervals from the stored query outcomes.

Assessable-domain analysis. For the revised joint-success presentation, an input-only mask removes specified non-carbon tetrahedral centers carrying a stereochemical label and ligand-internal metal atoms. It gives 55/15, 31/18, and 86/33 query/group counts for the initial, additional, and combined domains. We filter the query ledger before resampling, including for correctness strata. All contrasts within a cohort share 20,000 bootstrap draws (state 20272001 for the initial cohort and 20263001 for the additional and combined cohorts). Integer multiplicities accumulate the two-seed binary differences before division, retaining exact zero contrasts. Linear-interpolated 2.5th and 97.5th percentiles form the intervals. Intervals are recomputed after applying the input-domain mask. The script `code/analyze_assessable.py` records the mask, counts, group numbers, and all numerical settings. Its common-draw intervals are used for the original four-condition and selection-only assessable analyses. The composition-matched follow-up uses a separate paired group-bootstrap analysis with 20,000 draws per contrast. Its executable replay and manifest specify the contrast-specific random states and the ordered query/group inputs. These matched-control intervals do not use the common draws of the earlier four-condition analysis.

Historical transfer precedes its native-label audit. The second seed and nearest-donor comparison are follow-ups to the first-seed experiment. The follow-up analysis criterion requires a positive historical-minus-baseline effect at both seeds and a positive pooled lower confidence limit. Correctness strata and Top-k outcomes are secondary. The single reranking rule and group-deletion diagnostics were specified after the original outcomes were inspected and are post-hoc; their intervals are not multiplicity-adjusted.

A.5 Data and implementation availability

The anonymized `REPRO_SUPPLEMENT.zip` contains query/group assignments, all 800 pose banks as lossless float64 NumPy arrays with original ranks and confidences, the 32,000-pose endpoint ledger, historical-contract/reference provenance, separate input and native-evaluation graphs, receptor coordinates, and membership traces for 8,000 nearest-donor trajectories. Input-only selection runs separately from evaluation, with file access restricted to the historical inputs and baseline banks. The release includes recovered implementation snapshots and the executable evaluator/bootstrap function bodies.

```
python code/verify_package.py
python code/run_analysis.py
python code/analyze_assessable.py
```

The first command checks package hashes; the second regenerates the fixed-bank reranking, endpoint diagnostics, paired statistics, and group sensitivities without training or sampling. The third reproduces the assessable-domain tables and input-only exclusion ledger. The recorded analysis uses Python 3.11.15, NumPy 2.4.6, and RDKit 2025.09.6, pinned in `requirements-analysis.txt`. A separate CPU replay used Python 3.13.5, NumPy 2.3.5, and RDKit 2025.09.4 and reproduced the selected endpoints and four canonical intervals. The domain-restricted tables were generated in that verification environment; `requirements-verification.txt` and its manifest record those versions. Native labels are accessed only after selection. Baseline-bank availability is unchanged for every query and seed under the permutation.

The composition-matched extension is released in an additional `MATCHED_PRIMARY86_REPRO` directory. It contains 6,880 matched-control poses from 86 assessable queries and two fixed seeds, 13,760 intervention-step trace rows, the 33-group analysis roster, and per-query outcomes for all five conditions. Portable scripts reproduce the matched-control coordinate endpoints and the reported paired contrasts from these files. SDF headers and nonessential properties are redacted without changing atom order, coordinates, stored confidence, or rank; original and exported file hashes are provided.

For the original four-condition archive, original SDF byte files are not included, but parsed coordinates and their content hashes are preserved. The matched-control add-on instead supplies header-redacted SDF files whose coordinates and scoring metadata were verified against the original files. The original run manifests also do not pin every transitive source/model hash; recovered current dependency snapshots are distinguished from contemporaneously pinned wrappers. These limits affect regeneration of the original pipeline, not the supplied fixed-bank statistical replay.

B ADDITIONAL RESULTS FOR HISTORICAL GUIDANCE

B.1 ORIGINAL FULL-COHORT ANALYSES

The original analyses use all 62 initial and 100 combined queries, counting inputs outside the stereochemistry checker's scope as joint failures. Table 11 reports the original full-cohort paired effects and contrast-specific confidence intervals. These are the original analysis populations; the main text now uses the pooled 86-query assessable domain. Restricting the joint endpoint to the supported input domain changes the estimand and its group set, but does not add successful predictions. The reporting-stage mask and reduced-domain outcomes are given in Appendix A and the main text.

Table 11: Original full-cohort analysis, with unsupported inputs counted as joint failures. Entries are paired two-seed effects in percentage points and the original 95% intervals.

Comparison	Initial: 62 queries 17 groups	Combined: 100 queries 39 groups
Historical − no guidance	+12.10 [4.35, 23.17]	+13.00 [6.70, 21.25]
Nearest-K − no guidance	+10.48 [3.77, 19.49]	+9.50 [4.50, 15.28]
Native − no guidance	+17.74 [7.89, 30.23]	+16.50 [9.04, 24.77]
Historical − Nearest-K	+1.61 [−2.08, 8.57]	+3.50 [0.00, 8.79]
Donor reranking − no guidance	+8.06 [1.88, 18.52]	+7.00 [2.23, 13.58]
Historical − donor reranking	+4.03 [−3.41, 9.65]	+6.00 [1.10, 10.99]

Initial assessable cohort. Tables 12–13 retain the 55-query version of the comparisons.

Table 12: Joint Top-1 success on the 55 assessable queries in 15 groups from the initial 62-query cohort. Each condition produces 40 poses at each seed; the composition-matched condition was evaluated in a separate follow-up. The input-domain restriction is a reporting-stage analysis; the original full-cohort results are retained in Appendix B.1.

Condition	Seed 42	Seed 43	Mean success (%)
No guidance	10/55	13/55	20.91
Historical donors	19/55	19/55	34.55
Nearest-K (count only)	16/55	20/55	32.73
Composition-matched Nearest-K	16/55	19/55	31.82
Native-donor oracle	21/55	24/55	40.91

Table 13: Paired joint Top-1 effects on 55 assessable initial queries. Effects average the two seeds within query and intervals resample 15 groups. The upper block uses common draws from the original assessable reanalysis; the lower block gives the separately specified composition-matched follow-up. Differences are percentage points.

Comparison	Difference	95% confidence interval
Historical donors − no guidance	+13.64	[5.26, 26.67]
Nearest-K (count only) − no guidance	+11.82	[4.55, 22.13]
Native-donor oracle − no guidance	+20.00	[9.62, 33.67]
Historical donors − Nearest-K (count only)	+1.82	[−2.38, 9.76]
Composition-matched Nearest-K − no guidance	+10.91	[4.00, 21.05]
Historical donors − composition-matched Nearest-K	+2.73	[0.00, 10.17]

B.2 SEED-SPECIFIC PAIRED COMPARISONS

Table 14 gives the per-seed transitions and confidence intervals. An improvement is a query changing from failure to success relative to the stated comparator; a regression is the reverse. Summing these transitions across seeds counts query–seed outcomes, not distinct queries.

Table 14: Seed-specific Top-1 comparisons on 62 queries. Effects and group-cluster 95% confidence intervals are percentage points.

Seed	Comparison	Difference	Improvements/ regressions	95% interval
42	Historical − no guidance	+14.52	10 / 1	[3.77, 30.36]
43	Historical − no guidance	+9.68	11 / 5	[3.64, 18.37]
42	Nearest-K − no guidance	+9.68	7 / 1	[1.54, 20.29]
43	Nearest-K − no guidance	+11.29	9 / 2	[4.00, 20.45]
42	Native − no guidance	+17.74	12 / 1	[5.36, 35.59]
43	Native − no guidance	+17.74	13 / 2	[7.50, 27.12]
42	Historical − Nearest-K	+4.84	3 / 0	[0.00, 17.65]
43	Historical − Nearest-K	−1.61	4 / 5	[−4.92, 0.00]

B.3 RANKED-PREFIX SUCCESS

Table 15: Joint success among the first k confidence-ranked candidates in the original 62-query cohort. Each entry counts queries out of 62. All banks contain 40 poses; hence Success@40 equals availability for these conditions.

Seed	Condition	Top-1	Top-2	Top-4	Top-8	Top-40
42	No guidance	10	12	20	24	27
42	Historical donors	19	24	28	30	30
42	Nearest-K	16	24	30	34	34
42	Native-donor oracle	21	28	31	33	33
43	No guidance	13	19	24	26	27
43	Historical donors	19	23	26	27	29
43	Nearest-K	20	25	29	32	34
43	Native-donor oracle	24	28	31	33	35

Historical guidance has 47/124 successful query–seed banks at Top-2, compared with 31/124 for the baseline and 49/124 for Nearest-K. At Top-40, the corresponding totals are 59/124, 54/124, and 68/124. Native-donor guidance also has 68/124 Top-40 successes, although the individual successful queries need not be the same as those of Nearest-K. These prefix counts describe candidate production and ranking jointly.

At seed 42, the ten historical Top-1 improvements admit two different decompositions. Five selected trajectories become successful under correction and five were already successful but rise in rank. Separately, three improvements occur on queries with no successful pose anywhere in the baseline bank, and seven occur where a baseline success already exists. The partitions overlap and therefore are not additive.

Assessable correctness strata. Table 16 gives the corresponding initial-domain analysis for the original four conditions. These intervals use common draws within each represented stratum, independently of the composition-matched follow-up below.

Table 16: Retrospective contract-correctness strata among the 55 assessable initial queries. Each entry is a paired two-seed effect in percentage points with a 95% interval. Groups represented within each stratum are resampled; correctness is evaluated from native labels after transfer.

Comparison	Correct: 42 queries 11 groups	Incorrect: 13 queries 9 groups
Historical donors − no guidance	+21.43 [9.68, 39.06]	−11.54 [−19.23, 0.00]
Nearest-K − no guidance	+14.29 [5.41, 27.27]	+3.85 [0.00, 7.14]
Native-donor oracle − no guidance	+21.43 [9.68, 39.06]	+15.38 [0.00, 26.92]
Historical donors − Nearest-K	+7.14 [−2.56, 20.37]	−15.38 [−26.00, 0.00]

B.4 INCORRECT CONTRACTS AND TRAJECTORY-SPECIFIC DONOR SELECTION

Table 17: Same-cardinality historical errors in the original 62-query cohort: 12 queries, two seeds. Entries are pooled Top-1 effects in percentage points. Only the native-versus-baseline confidence interval is reported; other entries are descriptive point estimates.

Comparison	Difference
Historical − no guidance	−12.50
Nearest-K − no guidance	+4.17
Native − no guidance	+16.67 [5.00, 28.57]
Historical − Nearest-K	−16.67

Although both methods failed at Top-1 in all 17 incorrect-contract queries at seed 42, historical guidance removed baseline Top-40 support in four same-cardinality-error queries. With both seeds, the incorrect-contract Top-1 effect is negative. Correctness therefore affects more than whether a donor count is satisfied.

Nearest-K selects a native-exact set at step 18 in 548 trajectories at seed 42 and 564 at seed 43, out of 2,480 trajectories each. At each seed, 45 queries have at least one such trajectory. This trace is an observation about the generated candidates, not native information used by the control. The donor-count input remains the historical prediction, which can itself be wrong.

B.5 ADDITIONAL COHORT AND EXPANDED EXACTNESS ANALYSIS

The additional 38 queries share no Protein Data Bank entry or frozen group identifier with the initial 62. The additional-cohort counts in Table 18 are obtained by subtracting the primary-cohort counts from the combined 100-query counts and agree with the per-query records. The post-hoc intervals resample the additional cohort's 22 groups.

Table 18: Top-1 outcomes on the 38 additional queries in 22 groups. Mean rates average the two seeds.

Condition	Seed 42	Seed 43	Mean success (%)
No guidance	8/38	6/38	18.42
Historical donors	12/38	13/38	32.89
Nearest-K	10/38	10/38	26.32
Native-donor oracle	11/38	14/38	32.89

Table 19: Retrospective exactness strata in the expanded 100-query analysis. Effects average two seeds; intervals resample the groups represented within each correctness stratum (Appendix A.4). The exactness strata are not prospective eligibility rules.

Comparison	Correct sets ($n = 73$)	Incorrect sets ($n = 27$)
Historical − no guidance	+20.55 [11.93, 31.63]	−7.41 [−14.58, 0.00]
Historical − Nearest-K	+7.53 [1.33, 15.25]	−7.41 [−17.31, 2.78]

The added queries contain 28 correct and ten incorrect transferred sets, bringing the total from 45/62 to 73/100. The full expanded exact stratum has a positive historical-versus-nearest interval, whereas the original exact stratum's interval includes zero. The combined analysis adds observations but reuses the original cohort. Its conditional result does not replace the overall historical-versus-nearest comparison, whose interval still touches zero.

B.6 CONCENTRATION ACROSS METALS AND GROUPS

Seven of the 17 structural/site groups contribute a positive historical-versus-baseline effect, one a negative effect, and nine zero. Metal-specific denominators appear in Table 9; each query contributes

Table 20: Net Top-1 improvements by metal in the original 62-query cohort, summed over the two fixed seeds. These are query–seed differences in success counts, not per-metal success rates.

Metal	Historical − no guidance	Native − no guidance
Calcium	0	0
Cobalt	+7	+11
Iron	+2	+2
Magnesium	0	0
Manganese	+6	+9
Zinc	0	0
Total	+15	+22

two seed outcomes. Zero net change can include offsetting improvements and regressions. These counts therefore delimit the observed aggregate gain without establishing uniform benefit within any metal class.

B.7 Physical outcomes by evaluation domain

Table 21 gives the 55-query initial assessable domain, Table 22 gives the original 62-query cohort, and Table 23 gives all 100 collected queries, including the 14 input graphs outside joint-evaluation scope. Ligand and donor-identity predicates can still be examined on these inputs, and receptor clashes are measured directly from coordinates. The geometry-valid flag itself includes the stereo checker and therefore is false on unsupported inputs. Full-cohort and assessable-domain percentages answer different questions and are not pooled.

Table 21: Physical outcomes on the 55 assessable initial queries. Entries give seed 42, then seed 43. Selected-pose counts use 55 queries per seed; all-pose percentages use 2,200 candidates. The corresponding 62-query results are in Table 22; Table 23 instead reports the combined 100-query population.

Condition	Ligand-only Top-1 (n)	Clash at Top-1 (n)	All-pose clash (%)	All-pose valid (%)
No guidance	34, 39	4, 2	16.86, 15.64	83.14, 84.36
Historical donors	33, 36	6, 4	27.45, 26.00	72.55, 74.00
Nearest-K	36, 39	7, 7	26.32, 26.09	73.68, 73.91
Native-donor oracle	32, 38	6, 4	27.09, 27.09	72.91, 72.91

Table 22: Physical outcomes on all 62 initial queries, including unsupported stereochemical inputs. Entries give seed 42, then seed 43. Selected counts use 62 queries per seed; candidate fractions use 2,480 poses per seed.

Condition	Ligand Top-1	Top-1 clashes	All-pose clash (%)	Geometry-valid (%)
No guidance	37, 41	5, 4	18.51, 16.53	73.75, 74.84
Historical donors	34, 37	8, 8	30.20, 28.75	64.35, 65.65
Nearest-K (count only)	39, 41	10, 10	27.78, 27.34	65.36, 65.56
Native-donor oracle	33, 39	10, 7	29.80, 29.56	64.68, 64.68

B.8 Post-hoc preservation of baseline candidates

After inspecting the initial seed-42 results, we also considered a union of the 40 unmodified and 40 historically guided candidates, ranked by the original SurfDock confidence model. No new coordinates were generated. The diagnostic was defined after the primary result and uses twice as many stored candidates. Its success predicate was recorded as *Robust Full*, whereas the primary

Table 23: Full collected 100-query physical outcomes. Entries give seed 42, then seed 43. All-pose percentages use 4,000 poses per condition per seed. The separate geometry-valid flag includes supported stereochemistry.

Condition	Ligand	Identity	Top-1 clashes	All-pose clash (%)	All-pose valid (%)
No guidance	62, 62	36, 40	9, 6	17.77, 16.52	71.12, 71.92
Historical donors	57, 59	58, 60	15, 15	29.60, 29.85	62.67, 62.50
Nearest-K	63, 64	51, 51	14, 13	26.00, 26.52	63.62, 63.10
Native-donor oracle	55, 61	68, 71	17, 13	29.07, 29.32	63.00, 62.27

four-condition experiment explicitly uses the clash-free joint criterion. The union predicate has not been verified as identical to the main joint predicate.

Table 24: Seed-42 candidate-union diagnostic on 62 queries. Success entries retain the diagnostic's Robust Full definition. Success@40 is a ranked prefix of an 80-pose union, not its full-bank availability.

Outcome	No guidance (40 poses)	Historical (40 poses)	Union (80 poses)
Success@1	10	19	14
Success@40	27	30	34
Ligand-only Top-1	37	34	39
Clash at Top-1	5	8	4

Within this diagnostic, union minus baseline is +6.45 points [1.61, 13.95] at Top-1, with four improvements and no regressions; the Top-40 count rises by seven. The selected pose comes from the baseline bank in 45 queries and the guided bank in 17. Full-bank availability at 80 candidates was not tabulated separately. Set inclusion preserves every baseline candidate in the untruncated union, but it does not guarantee preservation in a confidence-ranked prefix or preservation of Top-1 correctness.

B.9 Selection-only comparator and component diagnostics

The single donor-aware permutation in Equation (7) was locked before its selected outcomes were computed, after the original historical-guidance results were known. All 200 baseline banks were reranked. Table 25 separates the initial, additional, and combined assessable analyses. The selector

Table 25: Post-hoc selection-only comparison within the stereochemistry checker's input domain. Counts are successes at seed 42 and seed 43. The combined sample contains the initial and additional cohorts. Differences are percentage points with group-bootstrap intervals.

Cohort	No guidance	Donor reranking	Historical donors
Initial (55)	10, 13	15, 18	19, 19
Additional (31)	8, 6	10, 8	12, 13
Combined (86)	18, 19	25, 26	31, 32

Cohort	Reranking − no guidance	Guidance − reranking
Initial (55)	+9.09 [2.17, 21.25]	+4.55 [−3.92, 10.81]
Additional (31)	+6.45 [−1.61, 15.91]	+11.29 [2.94, 21.43]
Combined (86)	+8.14 [2.73, 15.79]	+6.98 [1.28, 12.86]

reads only the initial ligand graph, supplied historical donors and metal, fixed target entries for those donors, and baseline coordinates/confidence. Native graphs, donor sets, native cardinality, native distances, RMSD, and endpoint labels are excluded from its input process. Whole-input-graph symmetry applies one chirality-preserving permutation to the complete supplied set. Unsupported graphs, empty or invalid contracts, and nonfinite banks would retain the complete saved ranking; none occurred.

The overlap deficit is the first sorting key, followed by descending frozen confidence and ascending sample identifier. Fixed-target radial residuals are recorded as diagnostics but never affect the ordering. The original confidence order uses descending NumPy sorting, and is retained exactly for the no-guidance baseline and any fallback. Every pose remains in the bank, so Success@40 is identical to the baseline on every query and seed. Selection changes on 99 of 200 banks. The input-access record, selected sample IDs, per-pose ranking keys, and paired comparison tables accompany the supplement.

Table 26: Component outcomes for each method's selected Top-1 pose in the original 62-query cohort. Counts are seed 42 / seed 43, out of 62 at each seed. Radial and directional marginals each allow a complete assignment but need not share a witness.

Condition	Ligand	Identity	Radial	Direction	Joint
No guidance	37/41	18/20	16/22	27/24	10/13
Donor reranking	39/40	37/38	24/28	34/31	15/18
Historical donors	34/37	35/35	33/36	23/24	19/19
Nearest-K	39/41	31/32	30/33	24/26	16/20
Native-donor oracle	33/39	39/42	37/42	25/30	21/24

All five conditions preserve supported stereochemistry on 55/62 queries at each seed and sanitize on 62/62. The reranked Top-1 has five fatal-clash cases at each seed; the other conditions' counts for the same 62-query denominator are given in Table 22. Higher identity accuracy alone is not sufficient for higher joint success: radial/directional geometry and whole-ligand placement must hold on the same selected pose under the joint definition.

We define two baseline-fixed diagnostic subsets separately at each seed. The first contains baseline ligand successes with incorrect donor identity; the second contains baseline ligand successes that fail the joint criterion. Every method is evaluated on those same baseline-defined queries. Table 27 reports recovered identity with ligand placement retained, joint recovery, and loss of ligand accuracy.

Table 27: Recovery and regression on baseline-fixed subsets. Entries are seed 42 / seed 43 counts. The denominator differs between the two panels. Identity recovery in the first panel is newly achieved; identity-plus-ligand in the second panel may already hold at baseline and is therefore not reported as a new identity recovery.

Baseline ligand-correct / identity-wrong: 24 / 23 queries

Condition	Identity repair with ligand retained	Joint recovery	Ligand loss
Donor reranking	11/14	2/6	3/4
Historical donors	10/13	6/7	8/4
Nearest-K	10/9	5/4	3/4
Native-donor oracle	13/16	8/8	9/4

Baseline ligand-correct / joint-wrong: 27 / 28 queries

Condition	Identity and ligand correct	Joint recovery	Ligand loss
No guidance	3/5	0/0	0/0
Donor reranking	14/19	2/6	3/4
Historical donors	13/17	7/9	8/5
Nearest-K	13/14	6/7	3/4
Native-donor oracle	16/20	9/10	9/5

B.10 Group deletion and donor-membership traces

Every frozen group is excluded once in the leave-one-group-out sensitivity analysis (Table 28). Removing the group containing the 16 consensus-extension queries leaves 46 initial queries, with historical-minus-baseline and historical-minus-nearest point differences of 14.13 and 2.17 points. These leave-one-group-out estimates are descriptive sensitivity checks.

Table 28: Ranges of leave-one-group-out paired point estimates in percentage points. They are not confidence intervals. Each row removes every group in the specified cohort once.

Cohort	Groups	Historical − no guidance	Historical − Nearest-K
Initial (62)	17	[9.32, 14.13]	[−0.86, 2.83]
Additional (38)	22	[10.29, 17.14]	[4.41, 8.57]
Combined (100)	39	[11.34, 14.29]	[2.08, 4.40]

The saved step-18 and step-19 memberships cover all 8,000 nearest-donor trajectories. All use the historical K; the traces confirm that every trajectory retains its chosen set at step 19, as required by the implementation. Table 29 compares these actual sets with the historical set using input-only graph symmetry. The donor-element multiset is not constrained and often differs. The controlled comparison consequently changes the atom set and sometimes its element composition, while holding K, update timing, operator, and sample count fixed.

Table 29: Nearest-donor membership from saved trajectories. Counts are seed 42 / seed 43. Set agreement uses whole-input-graph symmetry; element agreement compares multisets. Membership switching is zero throughout.

Cohort	Trajectories/seed	Historical set agreement	Element multiset agreement
Initial	2,480	580/599	1,527/1,563
Additional	1,520	395/401	1,055/1,060
Combined	4,000	975/1,000	2,582/2,623

B.11 Initial-cohort composition-matched comparison

This section retains the initial-cohort follow-up, distinct from the pooled 86-query comparison in the main text. The composition-matched control selects the nearest eligible atoms separately within each historical nitrogen, oxygen, and sulfur quota. It retains the selected set through both correction steps and all internal solver iterations. The same parent states, corresponding noise, radial targets, maximum correction iterations, and frozen confidence model are used as for historical guidance. All 124 query–seed banks completed with 40 poses; there were no missing or partial banks. Predictions, traces, technical results, and the evaluator were hash-frozen before endpoint evaluation. Technical checks covered one, two, and three oxygen donors, one sulfur donor, and a one-nitrogen/one-oxygen set.

The initially specified endpoint for this follow-up was historical-minus-composition-matched joint Top-1 success on the 55 inputs assessable by the stereochemistry checker. The 15 represented structural groups are the dependence units. The two seed differences are averaged within each query before resampling groups. Table 30 gives the reported paired intervals and query–seed transitions. The original study-wide historical-versus-baseline analysis and the full-cohort results remain separate.

Table 30: Composition-matched follow-up on 55 queries in 15 groups. Intervals are paired group-bootstrap 95% intervals; differences are percentage points. Transition counts describe 110 query–seed pairs, not 110 independent queries.

Comparison	Difference	95% interval	Improvements/ regressions/ties
Historical − composition-matched	+2.73	[0.00, 10.17]	7 / 4 / 99
Composition-matched − no guidance	+10.91	[4.00, 21.05]	15 / 3 / 92
Historical − no guidance	+13.64	[5.26, 26.39]	21 / 6 / 83

The historical-minus-matched difference is +5.45 points at seed 42 and zero at seed 43. Its pooled lower limit is zero, and the prespecified strict-superiority criterion is not satisfied. Including the seven unsupported inputs as automatic failures gives a 62-query sensitivity effect of +2.42 points [0.00,

8.82], with seven improvements, four regressions, and 113 ties. This sensitivity uses the original 17 groups and does not change the conclusion.

Component outcomes. Table 31 separates coordination, whole-ligand placement, and clashes. The coordination-witness indicator, named `anchor` in the evaluation code, requires joint donor-membership and geometric agreement under one mapping. It is distinct from the identity-only marginal in Table 26. Ordinary Full combines ligand success and the coordination witness; joint success further includes stereochemical and clash checks.

Table 31: Selected Top-1 components and candidate availability in the composition-matched follow-up. Counts pool two seeds over 55 queries, giving 110 query–seed observations. Differences are historical minus composition-matched, in percentage points. These are secondary intervals without multiplicity adjustment.

Endpoint	Historical	Composition-matched	Difference	95% interval
Coordination witness	42	39	+2.73	[−2.78, 9.62]
Ligand + coordination witness	38	36	+1.82	[−3.19, 9.23]
Joint success	38	35	+2.73	[0.00, 10.17]
Ligand-only success	69	72	−2.73	[−7.75, 5.00]
Fatal receptor clash	10	15	−4.55	[−8.57, −1.14]
Joint availability at Top-40	59	68	−8.18	[−14.71, 0.00]

Historical guidance has a lower selected-clash rate than the composition-matched control, but the ligand-only difference is uncertain. Across all 4,400 assessable candidates per condition, clash fractions are 16.25% for no guidance, 26.73% for historical guidance, 26.20% for cardinality-only Nearest-K, and 29.57% for composition-matched Nearest-K. Thus both nearest-donor controls and historical correction increase clashes relative to unmodified sampling.

Correctness and dependence. Table 32 conditions on the native-label audit of historical sets. All seven improvements occur among correct transferred sets, and all four regressions occur among incorrect sets. The retrospectively defined strata share site groups, and both intervals include zero.

Table 32: Retrospective historical-contract strata. Effects are historical minus composition-matched joint Top-1 success. Groups represented within each stratum are resampled. The strata share some groups.

Historical set	Queries	Groups	Difference [95% interval]	Improvements/ regressions
Correct	42	11	+8.33 [0.00, 20.83]	7 / 0
Incorrect	13	9	−15.38 [−26.19, 0.00]	0 / 4

Assignment comparison. The current 86-query assignment analysis in Section 5.3 uses the trajectory selected by the composition-matched condition's own frozen-confidence ranking. It evaluates that trajectory's assigned set against the native donor annotation, not against the historical condition's independently selected pose. Supplied-set agreement and final coordinate-based donor recovery are different quantities.

Group concentration and availability. Only two groups have discordant joint Top-1 outcomes. Group A has three improvements and three regressions; group B has four improvements and one regression. The remaining 13 groups contain only ties. The net three-outcome advantage comes entirely from group B. Group identifiers are retained in the supplementary aggregate table. At Top-40, historical guidance has no improvements and nine regressions relative to the composition-matched control. Greater candidate availability for the nearest-donor policy coexists with slightly higher selected success for historical guidance; neither result alone identifies a causal pathway through confidence or geometry.

Analysis release. The matched-control add-on supplies the saved banks, assignment traces, per-query outcomes, and bootstrap replay for the 86-query assessable domain. This includes 110 banks from the 55 assessable queries in the initial cohort and 62 banks from the 31 assessable additional queries. The seven unassessable queries in each cohort are outside this primary-domain release. The earlier all-62 aggregates remain a separate historical analysis record. The follow-up's historical-minus-baseline interval [5.26, 26.39] differs slightly from the earlier common-draw reanalysis [5.26, 26.67], despite identical counts. Each interval is kept with its respective analysis.

B.12 Pooled assignment and baseline-fixed diagnostics

The pooled follow-up analysis has 86 queries and 33 groups. Composition-matched nearest-donor guidance assigns the native set to 2,688 of 6,880 trajectories. Its own confidence-selected trajectory has an exact assigned set in 107 of 172 query–seed observations. In the diagnostic stratum compatible with the native donor count and elemental quotas, the corresponding denominator is 164 and exactness is 65.24%. Compatibility is determined with native labels only for analysis; all 172 observations remain in the policy comparison. Across both seeds, 78 queries have at least one exact assignment among their 80 trajectories. This is assignment coverage, not joint-success availability.

The 107 selected exact assignments contain 57 joint successes and 50 failures; the 65 incorrect assignments contain no joint successes. Assigned-set correctness and realized donor membership are not interchangeable: the operator constrains supplied donor distances but contains no non-donor exclusion term.

For the baseline-fixed diagnostic, let L^a_{is}, I^a_{is}, and J^a_{is} denote ligand, realized-identity, and joint success for condition a. We form two distinct sets from the unmodified baseline:

$$\mathcal{S}_{\text{joint}} = \{(i,s) : L^0_{is} = 1,\ J^0_{is} = 0\}, \tag{12}$$

$$\mathcal{S}_{\text{id}} = \{(i,s) : L^0_{is} = 1,\ I^0_{is} = 0\}. \tag{13}$$

On the assessable domain, $|\mathcal{S}_{\text{joint}}| = 69$: historical guidance recovers 26 joint successes and loses ligand accuracy on 14 observations. The narrower $\mathcal{S}_{\text{id}}$ contains 53 observations, with 18 joint recoveries and 12 ligand losses; 29 recover donor identity while retaining ligand accuracy. These quantities are recalculated from the same selected-pose ledger and input-domain mask. Their query–seed counts may include the same query twice, and neither subset is chosen using the intervention outcome.

Available granularity. The pooled composition-matched follow-up releases all 86-query seed-by-seed outcomes, all 172 matched coordinate banks, and both-step donor-assignment traces. The CPU replay reconstructs coordinate endpoints and the direct HIST–matched contrasts, including historical-correctness strata and group-wise paired wins and harms. The two seeds and 40 poses remain repeated measurements within each query; uncertainty is estimated by resampling the 33 predefined structural groups. The 55-query follow-up tables above therefore remain a separate cohort-specific record rather than a substitute for those pooled inputs. The existing stored-bank replay covers the original four sampling conditions and donor-aware reranking.

C Supporting studies of donor information and pose selection

The following experiments use separate populations and serve as supporting analyses. They are not pooled with the initial or expanded historical-transfer cohorts. The 800-bank replay package covers the historical-transfer experiments, not regeneration of these separate supporting studies.

C.1 Native-donor control on two docking backbones

A separate study evaluates all-40 last-two-step correction with native donor sets on 206 complexes. The second backbone is **DiffDock-L (v1.1 checkpoints)** (Corso et al., 2023; 2024). Its execution configuration specifies `model_dir: ./workdir/v1.1/score_model` and `old_score_model: false`; recovered generator and confidence checkpoint SHA-256 digests match the official v1.1 distribution. The model release is identified by the recovered configuration and checkpoint digests; contemporaneous container and loaded-weight hashes were not recorded.

Both SurfDock and DiffDock-L improve confidence-selected joint success (Table 33). Candidate availability increases by 5.34 points for SurfDock and 10.19 points for DiffDock-L; the latter has a paired 95% interval of [3.88, 16.50] points. Both backbones were evaluated with native donor sets.

Table 33: Separate 206-complex native-donor experiment with SurfDock and DiffDock-L. DiffDock-L uses the v1.1 generator and confidence checkpoints. Top-1 uses the frozen confidence model; availability is success among all 40 candidates. Intervals are paired 95% confidence intervals in percentage points.

Backbone	Baseline Top-1 (%)	Native donors Top-1 (%)	Difference (points)	95% interval
SurfDock	27.67	49.03	+21.36	[13.59, 29.13]
DiffDock-L	14.56	20.39	+5.83	[0.97, 10.68]

Availability: SurfDock 67.96% → 73.30%; DiffDock-L 34.47% → 44.66%.

C.2 Learned donor sets and downstream effects

An earlier experiment uses a structured ligand-graph and metal-element donor predictor with approximately 4.87 million parameters. On 104 queries and two seeds, its output is passed to the same class of last-two-step distance correction. The mean selected joint-success rates are 25.96% without guidance, 25.00% with predicted donors, and 41.35% with native donors. The native-versus-baseline difference is +15.38 points [3.75, 24.04].

Retrospective exactness separates opposing effects. On 102 query–seed banks with correct predicted sets, baseline success is 33.33% and predicted-donor success is 46.08%, a +12.75-point change. On 106 banks with incorrect sets, success changes from 18.87% to 4.72%, a -14.15-point change. Thus the donor predictor's average accuracy does not by itself describe the balance of improvement and degradation under intervention. These 208 banks are repeated evaluations of 104 queries, not 208 distinct biological samples.

C.3 External pretraining of a donor predictor

A separate donor-prediction comparison uses the same structured architecture and target evaluation while varying ligand-encoder pretraining. The target-only baseline reaches 35.93% complete-set exactness, a matched externally pretrained model 37.27%, and a variant using a listwise complete-set pretraining objective 36.37%. The listwise variant differs from the matched pretrained model by -0.90 points, with a component-bootstrap interval of $[-2.04, 0.34]$. Same-cardinality wrong-set errors rise from 29.78% to 31.65%.

The external corpus contains 70,069 records, of which 69,997 pass label and atom-mapping checks. Compatibility filtering gives 41,227 records, and exact-ligand/scaffold exclusions against the target collections leave 33,837. The listwise pretraining variant did not improve complete-set prediction over the matched pretrained baseline.

C.4 Historical-only evidence screening

We also evaluated a stricter eligibility rule using only earlier reference records. Each historical record is treated as a pseudo-query, with transfer restricted to still earlier structures; canonical ligand graphs are deduplicated. A family qualifies only if it has at least three distinct historical graph units and no observed transfer errors. There are 1,146 pseudo-query trials, of which 369 are assessable, and no current query qualifies.

Among the 33 queries covered by strict site-matched transfer, twelve have no assessable historical graph and twelve have only one. No zero-error family has two or more independent evidence units. Families with four and nine evidence units have historical precision of 75% and 22.22%, respectively. The historical-only screen is distinct from the broader transfer rule used in the docking experiments. The rule's zero coverage reflects limited consistent historical evidence.

C.5 CHEMICAL-DESCRIPTOR MODELS FOR COMPLETE DONOR SETS

Three classical learning methods test whether chemical and pocket descriptors improve the frozen MetalloDock donor-head output. The shared development data contain 339 queries in 172 biological clusters and 727,650 candidate sets of eligible nitrogen, oxygen, and sulfur atoms, with cardinality one to four. Five outer folds and fixed inner splits are reused. Labels use whole-graph chirality-preserving donor-set equivalence, and only out-of-fold predictions enter the result tables.

The primary metric is cluster-macro exactness: within-cluster query accuracy is averaged equally over 172 clusters. The adjacent case counts instead weight all 339 queries equally. Paired percentile intervals resample the 172 cluster means 10,000 times with NumPy RandomState seed 20260922. These settings belong to this development study, not to the historical replay.

Shared descriptors describe atom chemistry, selected donor composition, pair topology, metal-centered protein environments, pocket–set interactions, and Gasteiger partial-charge summaries (Gasteiger & Marsili, 1980). Retained water information is absent in all 339 inputs. Gasteiger calculation fails on 14 queries, represented with missing-value flags rather than zero charge. The receptors are ligand-removed holo structures; pretrained-model exposure and protein-family independence are unresolved.

Models. The CatBoost (Prokhorenkova et al., 2018) hybrid uses a depth-four PairLogit ranker with 333 columns, including MetalloDock scores and outputs of the learned complete-set predictor. Training uses at most 500 iterations, learning rate 0.03, L2 coefficient 10, and normalized discounted cumulative gain at rank two for early stopping with patience 50. The log-linear model adds a 153-feature structured score to a fixed, train-standardized MetalloDock independence offset. It uses cluster-weighted catalogue cross-entropy with L2 coefficients in $\{0.1, 1, 10\}$. The kernel model adds a Ridge residual with a 128-component Nyström radial-basis approximation. Its charge-augmented version uses 322 features; the charge-free version uses 316, with training-only imputation and scaling. Inner cluster-macro Top-1 chooses regularization for both models.

Table 34: Prespecified chemical-descriptor comparisons against the MetalloDock expected-count decoder. Differences and 95% intervals refer to cluster-macro exactness.

Method	Cluster exact (%)	Exact queries (n/N)	Difference (points)	95% interval
MetalloDock expected-count	45.49	150/339	—	—
CatBoost hybrid	45.31	140/339	-0.18	$[-6.87, 6.56]$
Log-linear full model	46.21	125/339	+0.72	$[-5.87, 7.13]$
Charge-augmented kernel	41.44	114/339	-4.05	$[-8.50, 0.30]$

All three primary intervals include zero (Table 34). In the log-linear model, adding donor-pair features improves its unary/cardinality variant by +7.25 points [1.54, 13.13], while adding pocket interaction gives +0.57 $[-4.14, 5.34]$. The pair block also changes the cardinality representation through redundant count-derived features. Adding the Gasteiger block to the kernel changes exactness by -0.58 $[-3.49, 2.33]$ points; it also changes the kernel distance and bandwidth. Removing added pocket descriptors leaves any protein information already encoded in the frozen predictor.

Compared with MetalloDock expected-count, CatBoost gains 37 correct queries and loses 47; log-linear gains 35 and loses 60; the kernel gains five and loses 41. These earlier chemical-descriptor runs use RDKit 2023.03.1 (Landrum et al., 2023); the current saved-bank analysis uses release 2025.09.6. A post-hoc restriction to the MetalloDock-predicted donor count raises the log-linear pair/full variants to 47.26/47.73% cluster-macro exactness, with 150/339 and 152/339 correct queries. The respective baseline differences are +1.77 $[-2.83, 6.29]$ and +2.24 $[-1.50, 6.13]$ points. These exploratory decodings were not used for the historical docking experiments. The model comparison jointly varies the learner, prior usage, feature representation, and training loss.

C.6 SELECTION FROM MIXED CANDIDATE BANKS

An exploratory study trains a metal-local ranker on a separate 192-complex development collection and evaluates it on 104 queries. It combines 40 unmodified and 40 predicted-donor-guided poses,

retains the common-confidence top eight, and chooses one with a learned metal-local ranker. The feature set includes confidence, rank, contact/clash and stereochemistry features, plus metal-distance, cardinality, and overcoordination summaries. Common confidence on the mixed bank gives 27.40% joint success, the ranker on an uncorrected bank 35.58%, and the ranker on the mixed bank 36.06%. The mixed-versus-common-confidence difference is +8.65 points [1.49, 13.54], but the difference from the uncorrected-bank ranker is only +0.48 points. This supports the usefulness of metal-local selection features without isolating a benefit from adding corrected candidates. The main historical-transfer experiment keeps the original confidence model fixed.

D EARLIER CONTROLLER CONFIGURATION

Before the last-two-step radial correction, we evaluated a different controller that branches from a step-five SurfDock state and applies corrections after steps 6, 8, 10, 12, 14, 18, and 19. This controller uses one damped Gauss–Newton update per call. Its residual combines donor distance targets, movement from the incoming pose, and non-donor ordering. With $d_i = \|\mathbf{x}_i - \mathbf{z}_m\|_2$, its ordering term is proportional to

$$\max\left\{0,\ \max_{i\in D} d_i + 0.20\,\text{Å} - \min_{j\notin D} d_j\right\}. \tag{14}$$

Non-members include all ligand heavy atoms outside the supplied set. The linear step solves

$$(J^\top J + 0.10I)\Delta q = -J^\top r. \tag{15}$$

It uses separate translation and angular caps in addition to a Cartesian RMS cap. A trial that increases the count of protein contacts below the specified threshold is rejected. This differs from the radial-only last-two-step operator, which has no contact guard, movement residual, or exclusion residual.

Earlier branch evaluations use a shared bank of 16 unmodified poses together with 16 additional poses from a correction condition. Their primary endpoint measures whether at least one successful pose occurs in the union. The main experiment instead applies a last-two-step operator to all 40 trajectories and evaluates the confidence-selected Top-1 pose. Preserving an original bank changes the possibility of losing baseline support, while selecting the best-ranked pose asks a different question from availability. These designs are therefore kept separate when interpreting the method.

In a 50-case scheduling comparison, the unmodified condition has eleven successes, fixed membership and an open-loop variant each have 23, and two adaptive variants have 20 and 22. Fixed membership matches the highest observed success count in this 50-case comparison. Several solver, timing, and penalty settings differ across the controller configurations.

E CONTRACT ACCURACY AND INTERVENTION BENEFIT

For a fixed population, operator, selector, and endpoint, let $A = 1$ indicate a supplied contract, $Z = 1[\widehat{D} \equiv D^\star]$ indicate native-set agreement, and d_i be query i's two-seed paired Top-1 change. Define $p = \mathbb{P}(Z = 1 \mid A = 1)$, $g = \mathbb{E}[d \mid Z = 1, A = 1]$, and $\ell = \mathbb{E}[d \mid Z = 0, A = 1]$. The law of total expectation gives

$$\Delta = pg + (1-p)\ell. \tag{16}$$

The conditional terms are net changes, and their weighting must match the overall estimand. For the main assessable population,

$$\widehat{\Delta}^{86}_{\text{historical,baseline}} = \frac{69}{86}\frac{30}{138} + \frac{17}{86}\frac{-4}{34} = \frac{26}{172} \approx 0.1512. \tag{17}$$

The retained earlier analyses give

$$\widehat{\Delta}^{62}_{\text{historical,baseline}} = \frac{45}{62}\frac{18}{90} + \frac{17}{62}\frac{-3}{34} = \frac{15}{124}, \tag{18}$$

$$\widehat{\Delta}^{62}_{\text{historical,nearest}} = \frac{45}{62}\frac{6}{90} + \frac{17}{62}\frac{-4}{34} = \frac{2}{124}, \tag{19}$$

$$\widehat{\Delta}^{100}_{\text{historical,nearest}} = \frac{73}{100}\frac{11}{146} + \frac{27}{100}\frac{-4}{54} = \frac{7}{200}, \tag{20}$$

$$\widehat{\Delta}^{55}_{\text{historical,composition}} = \frac{42}{55}\frac{7}{84} + \frac{13}{55}\frac{-4}{26} = \frac{3}{110}. \tag{21}$$

These identities decompose the observed paired effects by contract correctness. The 55-query composition contrast is distinct from the pooled follow-up 86-query comparison.

For expected incremental cost $c \geq 0$ in the same units, $U = pg + (1-p)\ell - c$. When $\ell = -h \leq 0$ and $g + h > 0$, positive utility requires $p > (h + c)/(g + h)$. This threshold depends on the supplier's actual error distribution. Per-pose clash fractions and per-query success cannot be subtracted without specifying their common utility scale.

If simultaneous bounds give $p \geq p_L$, $g \geq g_L$, $h \leq h_U$, and $c \leq c_U$, with $g_L + h_U \geq 0$, then

$$U \geq p_L g_L - (1 - p_L) h_U - c_U. \tag{22}$$

Indeed, substitution bounds U by $p(g_L + h_U) - h_U - c_U$, which is nondecreasing in p. Separate marginal 95% intervals do not imply the required simultaneous event. Using the bound as a selection rule requires separate calibration and validation (Bates et al., 2021; Angelopoulos et al., 2025).

For the original roster only, unchanged baseline outputs outside the 62 supplied queries would imply $(62/304)(15/124) = 15/608 \approx 0.0247$. This accounting identity does not estimate the pooled 86-query policy's population coverage or provide a full-roster confidence interval.

F MOLECULAR EXAMPLES

Figure 2 illustrates a successful change and a regression in the initial cohort. Examples were selected retrospectively, taking the smallest ligand by heavy-atom count and then query identifier within each of two seed-43 categories: a joint-success improvement with a correct historical set, and a regression with an incorrect set of the correct cardinality. The rule selects small ligands for visual clarity. The plotted predictions are each method's selected Top-1 pose; they need not be the same trajectory index.

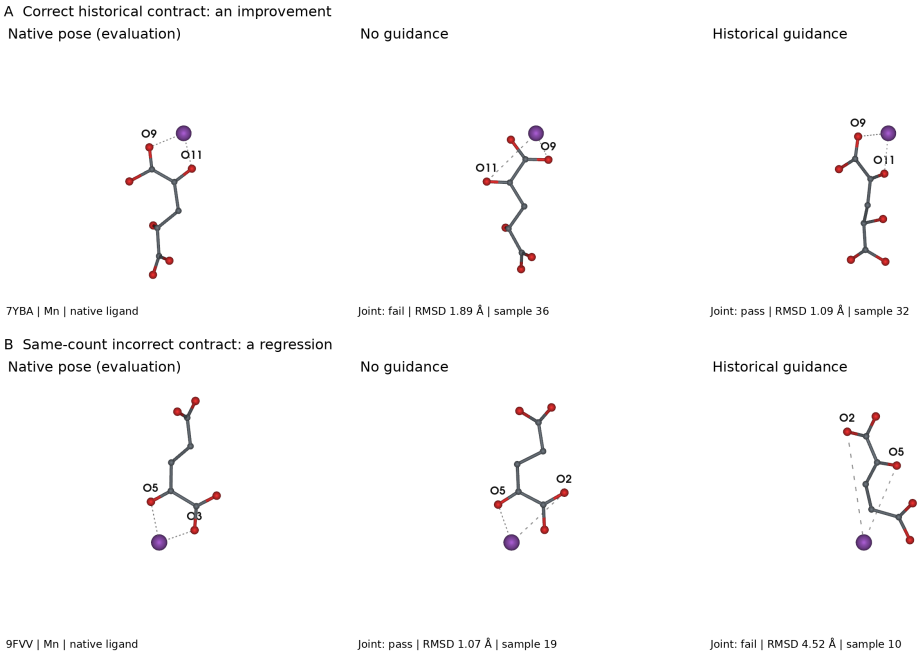


Figure 2: Saved molecular structures in a common receptor coordinate frame. (A) PDB 7YBA: a correct historical set accompanies a change from joint failure to success. (B) PDB 9FVV: the transferred set has the correct cardinality but different atom membership, and the selected pose loses joint success. Native poses are shown only for evaluation. Carbon is gray, oxygen red, and the target manganese ion purple; receptor atoms and hydrogen atoms are omitted for clarity. Labels give the element and one-based atom index. Dashed links identify the native donor set in the native panel and the supplied historical set in both prediction panels; these connectors label the specified donor sets. Cameras are fixed within each row, and structures retain their saved coordinates. Clashes are assessed quantitatively using the receptor coordinates (Section 4).

The supplementary renderer reads the stored coordinate arrays and native graphs directly. It records the query identifiers, selected samples, donor indices, camera parameters, and input hashes in `molecular_examples/FIGURE_PROVENANCE.json`. The plotted coordinates are exported without modification.